DOCUMENT RESUME

ED 304 462 TM 012 834

AUTHOR Mazzeo, John; Harvey, Anne L.
TITLE The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests: A Review of the Literature. College Board Report No. 88-8.
INSTITUTION College Entrance Examination Board, New York, N.Y.; Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-21
PUB DATE 88
NOTE 33p.
AVAILABLE FROM College Board Publications, P.O. Box 886, New York, NY 10101-0886 ($6.00).
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Comparative Analysis; *Computer Assisted Testing; Difficulty Level; *Educational Testing; Elementary Secondary Education; *Equated Scores; Higher Education; Literature Reviews; *Psychological Testing; *Psychometrics; Reading Tests; Standardized Tests
IDENTIFIERS *Paper and Pencil Tests

ABSTRACT

A literature review was conducted to determine the current state of knowledge concerning the effects of computer administration of standardized educational and psychological tests on the psychometric properties of these instruments. Students were grouped according to a number of factors relevant to the administration of tests by computer. Based on the studies reviewed, it seems that: (1) the rate at which examinees omit items in an automated test may differ from the rate at which they omit items in a conventional presentation; (2) scores on automated personality inventories are lower than scores obtained using the conventional testing format; (3) scores from automated versions of speed tests are not likely to be comparable with scores on paper-and-pencil versions; (4) presentation of graphics in an automated test may affect score equivalence; (5) tests containing items based on reading passages can become more difficult when presented via computer; and (6) the possibility of asymmetric practice effects may make it wise to avoid equating studies based on single-group counterbalanced designs. Ten data tables are provided. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED304462

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

PATRICIA K. HENDEL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

College Board Report

No. 88-8

The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests

A Review of the Literature

John Mazzeo
Anne L. Harvey

The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests

A Review of the Literature

John Mazzeo
Anne L. Harvey

College Board Report No. 88-8
ETS RR No. 88-21

College Entrance Examination Board, New York, 1988

John Mazzeo is a measurement statistician at Educational Testing Service, Princeton, New Jersey.

Anne L. Harvey is an associate measurement statistician at Educational Testing Service, Princeton, New Jersey.

Acknowledgments

Funding for the initial preparation of this report was provided by the College Board/ETS Joint Staff Research and Development Committee. Additional funds for revisions were provided by the College Board Division at ETS.

The authors would like to thank Isaac Bejar, Daniel Eignor, Phillip Oltman, and Spencer Swinton of ETS for their reviews of various versions of this manuscript.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101-0886. The price is $6.

Copyright © 1988 by College Entrance Examination Board. All rights reserved. College Board, Scholastic Aptitude Test, SAT, College Scholarship Service, CSS, FAF, and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Abstract 1

Introduction 1

Automated Testing Using Slides and/or Free-Response Items 2

Computer-Administered Tests 7
Tests Consisting of Single-Screen Items 7
Tests Containing Items Based on Reading Passages 23

Conclusions 25

References 26

Tables

1. Summary of Studies Using Nonstandard Equipment and/or Tests Containing Free-Response Items 3

2. Summary of Studies Using Untimed Personality Inventories or Psychological Scales 9

3. Means for Paper-and-Pencil and Computer Versions of the MMPI 10

4. Effect Sizes for Computer and Paper-and-Pencil Forms of the MMPI 10

5. Effect Sizes for Computer and Paper-and-Pencil Forms of the CPI 12

6. Summary of Studies Using Untimed Aptitude Tests 14

7. Summary of Studies Using Untimed Achievement Tests 16

8. Summary of Studies Using Timed Power Tests 17

9. Summary of Studies Using Speed Tests 19

10. Summary of Studies Using Tests Consisting of Single-Screen Items with Graphics 24

11. Summary of Studies Using Tests Containing Items Based on Reading Passages 24

ABSTRACT

A literature review was conducted to determine the current state of knowledge concerning the effects of the computer administration of standardized educational and psychological tests on the psychometric properties of these instruments. Studies were grouped according to a number of factors relevant to the administration of tests by computer. Based on the studies reviewed, we arrived at the following conclusions:

1. The rate at which test-takers omit items in an automated test may differ from the rate at which they omit items in a conventional presentation.
2. Scores on tests from automated versions of personality inventories such as the Minnesota Multiphasic Personality Inventory are lower than scores obtained in the conventional testing format. These differences may result in part from differing omit rates, as described above, but some of the differences may be caused by other factors.
3. Scores from automated versions of speed tests are not likely to be comparable with scores from paper-and-pencil versions.
4. The presentation of graphics in an automated test may have an effect on score equivalence. Such effects were obtained in studies using the Hidden Figures Test. However, in studies with three Armed Services Vocational Aptitude Battery (ASVAB) tests, effects were *not* found.
5. Tests containing items based on reading passages can become more difficult when presented on a CRT. This was demonstrated in a single study with the ASVAB tests.
6. The possibility of such asymmetric practice effects may make it wise to avoid conducting equating studies based on single-group counterbalanced designs.

INTRODUCTION

The automated administration of educational and psychological tests has been recognized for some time as an attractive alternative to the conventional paper-and-pencil or face-to-face modes of administration (Elwood 1969; Denner 1977; Space 1981). In an automated test administration the test-taker is presented items on a display device—for example, a slide projector, teletypewriter (TTY), or cathode-ray tube (CRT)—and then indicates his or her answers on a response device such as a standard keyboard, specially constructed response box, or touch-sensitive screen. The presentation of test items and the recording of test-taker responses are controlled by a mainframe or a minicomputer with time-sharing capabilities or by a microcomputer. The advantages of automated administration typically cited are increased test security; reduced costs of test production, administration, and scoring; and the possibility of immediate feedback of results.

Several alternative automated testing procedures have been suggested. One alternative that has received considerable attention is computerized adaptive testing (CAT). Unlike conventional testing, in which all test-takers receive a common set of items, adaptive testing is designed so that each test-taker receives a set of items with psychometric characteristics appropriate to his or her estimated level of ability. Most proposed systems for developing adaptive tests are based on Item Response Theory (IRT), since Classical Test Theory is not well suited to the task. Many psychometric issues need to be addressed and resolved before CAT tests are routinely used (Green et al. 1984).

A second alternative receiving increased attention is the automated administration of existing standardized tests that have been developed and normed in a paper-and-pencil or a face-to-face format. In this procedure the entire set of items from an existing test is administered to all test-takers by computer rather than by the paper-and-pencil or the face-to-face procedures associated with the standardized instrument. Most standardized tests consist of limited-choice item formats such as multiple-choice, true/false, and agree/disagree/don't know. However, some standardized tests—individually administered intelligence tests in particular—contain one or more subtests consisting of free-response items, responses to which must be evaluated by scoring rules provided by the test publisher.

Research has been published on the use of automated versions of (1) standardized intelligence tests such as the Wechsler Adult Intelligence Scale (WAIS) (Elwood 1972a, 1972b, 1972c, 1972d), the Slossen Intelligence Test (Hedl, O'Neil, and Hansen 1971), and the Raven's Standard Progressive Matrices (Calvert and Waterfall, 1982); (2) standardized personality measures such as the Minnesota Multiphasic Personality Inventory (MMPI) (Dunn, Lushene, and O'Neil 1972), the California Psychological Inventory (CPI) (Scissons 1976), and the Speilberger State-Trait Anxiety Inventory (Lukin et al. 1985); and (3) standardized aptitude tests such as the Armed Services Vocational Aptitude Battery (ASVAB) (Kiely, Zara, and Weiss 1986) and the Aptitude Classification Battery of the German Federal Armed Forces (Wildgrube 1982). With the increasing availability of microcomputers in mental health clinics, elementary schools, high schools, and colleges, a market for computer-administered versions of paper-and-pencil standardized tests is developing and will continue to grow.

As automated versions of standardized tests become available, users will have a choice of administering the tests by computer or by paper and pencil, which-

ever format is more convenient. A critical issue in the use of automated versions is the equivalence of computer-generated scores to corresponding paper-and-pencil scores. Can a score from an automated test be interpreted in the same manner as a score of equal magnitude from a conventional version?

The concern for score equivalence is explicitly acknowledged in the recently issued *Guidelines for Computer-Based Tests and Interpretations* (1986), developed jointly by the American Psychological Association's Committee on Professional Standards and the Committee on Psychological Tests and Assessment. Guideline 16 states:

> When interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cutting scores obtained from conventional tests. Scores from conventional and computer administrations may be considered equivalent when (*a*) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (*b*) the means, dispersions, and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode. (p. 14)

The concern over the comparability of standardized test scores from computer and conventional administration is, no doubt, well founded. In a recent article in the *Journal of Educational Measurement*, Green et al. (1984) suggest some possible ways that the psychometric characteristics of tests might be altered when items are switched from paper-and-pencil to computer administration. Tests may become more or less speeded, resulting in an overall shift in the difficulty of the test. Reading comprehension items, for example, might be more or less difficult when presented on a CRT because of the difficulties or advantages that emerge from the computer presentation of reading passages. More generally, the nature of the test-taking task might change so that those students who are more familiar with computers may perform somewhat better on the computer-administered version of the examination than do equally able students who are not so familiar. As a result, the tester may unintentionally be measuring computer literacy along with the subject matter.

Our report reviews the research studies that have compared the psychometric properties of automated and conventional test forms. This report does not concern itself with studies comparing computerized-adaptive testing with conventional paper-and-pencil testing. The intent is to identify what, if any, generalizations can be drawn regarding the factors that affect the comparability of scores from the two formats. Of principal interest have been test structure (discrete items versus sets of items based on a common reading passage or problem description), item content (items containing graphics versus items containing only verbal material), and test timing (speeded versus untimed tests), which writers like Green et al. have suggested might affect the psychometric properties of an automated version of an existing test.

The manner in which automated tests are administered has changed considerably over the last 20 years as the technology for delivering them has evolved. Our main concern has been to identify generalizations, which are important given the way in which automated test administration is currently conceived. With today's technology, the most typical scenario involves a microcomputer (or a minicomputer with time-sharing capabilities), a CRT display device, and a computer keyboard response device to administer tests consisting of limited-choice items. Thus, the greater portion of this paper reviews studies that have used automated testing procedures—similar to this current conception—to administer tests of limited-choice items. A number of studies were encountered, however, which used automated testing procedures based on the technology available at the time the studies were performed or used tests containing at least some free-response items. While these studies are less relevant than the previously mentioned group, we still consider their results important. Consequently, they are also reviewed, albeit in somewhat less detail.

The paper is organized into three major sections: The first section reviews the studies that used automated testing procedures that differ somewhat from the current conception. The second section details the studies that used technology similar to that available today; this section is divided into several subsections, each of which deals with tests containing particular types of limited-choice items. The final section summarizes the major conclusions.

AUTOMATED TESTING USING SLIDES AND/OR FREE-RESPONSE ITEMS

Many of the earlier studies of automated tests employed administration procedures that differ somewhat from the current conception of an automated test. Several of these studies were concerned with automated versions of tests (such as Wechsler Adult Intelligence Scale and the Slossen Intelligence Test) that have free-response formats. In addition, many of these studies used equipment such as a slide projector (or, in one case, a normal test booklet) rather than a CRT (or TTY) to present items. Varied response devices, from conventional keyboards to specially adapted boxes with buttons for each possible response, were used. The test administration control was usually handled by a specially designed mainframe computer program; however, in some cases an examiner was required to control the presentation equipment.

A summary of the 10 studies reviewed in this section is given in Table 1. For each study a number of

Table 1. Summary of Studies Using Nonstandard Equipment and/or Tests Containing Free-Response Items

				Experimental Design				*Results*	
					Controlled for:				
Study	*Test*	*Equipment*	*Type*	*Mode Factor*	*Order*	*Form*	*Subjects*	*Significant Differences*	*Statistics Affected*
Hedl, O'Neil, & Hansen (1971)	Slossen Intelligence Test	CRT/keyboard	Untimed	W*	Y‡	Y	—	Y	Mean
Knights, Richardson, & McNarry (1973)	Peabody Picture Vocabulary Test Coloured Progressive Matrices	Slides/touch-sensitive screen/audio disk Keyboard	Untimed	W	Y	Y	—	Y	Mean
Hitti, Riffer, & Stuckless (1971)	Raven's Standard Progressive Matrices	CRT/paper test booklets	Untimed	B†	—	—	Y	N	—
Overton & Scott (1972)	Peabody Picture Vocabulary Test	Slides/special response panel	Untimed	W	Y	Y	—	N	—
Klinge & Rodziewicz (1976)	Peabody Picture Vocabulary Test	Slides/special answer sheets	Untimed	W	Y	Y	—	N	—
Elwood (1972b)	WAIS	Special testing booth/slides/special keyboard	Untimed	W	Y	Y	—	Y**	Mean
Elwood (1972d)	WAIS	Special testing booth/slides/special keyboard	Untimed	W	N§	Y	—	N**	—
Calvert & Waterfall (1982)	Raven's Standard Progressive Matrices	Slides/special keyboard	Untimed	B	—	—	Y	N	—
Hoffman & Lundberg (1976)	Pathology final exam	Slides/response stations	Timed/power	W	Y	Y	—	Y	Mean
Curtis & Kropp (1961)	School and College Ability Tests	Slides/paper answer sheets	Timed/power	W	N	Y	—	Y	Mean/number of omits

*W = within-subject variable.

†B = between-subject variable.

‡Y = yes

§N = no

**Significance tests were not performed by the authors of the original study. However, sufficient data were reported to allow for the calculation of *t*-tests based on dependent samples.

items of information are included. Because similar formats are used in later tables, Table 1 is described here in detail.

The first column lists the authors and the publication date of each study. The next two columns describe features of the testing. Column two lists the test or tests that were administered. In column three the equipment used for the automated administration is briefly described. The third column is omitted from later tables because the studies reviewed in subsequent sections are relatively homogeneous with respect to equipment.

Column four lists the timing conditions. We distinguish among three types: (1) tests administered without a time limit (untimed); (2) tests administered with a time limit but primarily designed as power tests (timed power); and (3) tests administered with a time limit but primarily designed as speed tests. It is not unreasonable to expect that timed power tests and speed tests might be more sensitive to mode-of-administration effects. Any distractions or difficulties introduced by automated test administration, or advantages introduced by ease of responding, will affect the rate at which items are answered. Such rate differences could, in turn, have an effect on score comparability when tests are timed. In several studies the timing conditions associated with the tests were not explicitly stated and had to be inferred, to the extent possible, from the available information or from additional background sources describing the tests.

To show the degree of methodological soundness, we have listed features of the experimental design for each study in the fifth, sixth, seventh, and eighth columns. In column five, titled Mode Factor, we have indicated whether the mode of test administration—that is, automated presentation or conventional presentation—was a between-subject (B) variable (different subjects take the automated and the conventional versions of the test) or a within-subject (W) variable (the same subjects take the test in both versions).

Results pertaining to mode-of-administration effects from studies using within-subject designs could be confounded by the order of test administration (if one mode occurs first for all subjects). Results could also be confounded by test form differences (if different, nonequivalent forms of a test are administered in each of the presentation modes). Therefore, for each study using a within-subject (W) design, we have shown whether the experimental design controlled for possible order effects (in the column titled Order) and for effects due to test form difficulty (in the column titled Form). Form effects were usually controlled for either by counterbalancing test forms across modes of administration, employing parallel forms in the two test modes, or by using the same test form in each administration. Order effects were controlled for by counterbalancing.

Results pertaining to mode-of-administration effects from studies using a between-subject design could be confounded with intergroup differences in the constructs being measured. Therefore, for each study using a between-subject (B) design, we have indicated whether extraneous group differences were controlled for, in the column titled Subjects. Such controls consisted of random assignment or stratified random assignment of matched pairs to the experimental conditions.

The last two columns briefly describe the results in each study. Column nine shows whether any statistically significant effects (that is, main effects or interactions) were obtained. In most cases we have used the significance results reported by the authors of the studies. In some cases the original authors did not perform significance tests but provided the necessary summary statistics to assess statistical significance. In these instances, we calculated the necessary statistics and report the results. Such tests are indicated with a double asterisk in Table 1. When the results of significance tests were not reported and the necessary summary statistics were unavailable, we have used a dash to indicate "not applicable." Column ten indicates which score statistics (mean, standard deviation) differed significantly across the test administration modes. Further details concerning both the methodology and the results of each study are provided in our text discussion.

Significant mode-of-administration effects were obtained in five of the ten studies listed in Table 1—three of the eight studies involving untimed tests and both of the studies involving timed power tests. In all five studies (Hedl, O'Neil, and Hansen 1971; Knights, Richardson, and McNarry 1973; Elwood 1972b; Hoffman and Lundberg 1976; and Curtis and Kropp 1961), differences were observed in the mean scores obtained under automated versus conventional test administration procedures. In the Curtis and Kropp study, statistically significant differences were also observed in the mean number of omitted items; however, in that study the order of test administration was perfectly confounded with the mode of administration, making the results somewhat difficult to interpret.

Beyond the issue of statistical significance, an interesting pattern of results is present across the Table 1 studies, considered as a set. With the exception of Curtis and Kropp (1961) and Hitti, Riffer, and Stuckless (1971), the score means produced from the automated version of each test were lower than the mean scores from the conventional version. The mean differences are seldom large and, as mentioned, not always statistically significant. However, mean scores from the automated versions were consistently lower although the studies used various types of automated equipment and different tests.

Klinge and Rodziewicz (1976), Knights, Richardson, and McNarry (1973), and Overton and Scott

(1972) all studied the effects of automated administration on the Peabody Picture Vocabulary Test. Very little information is presented on the exact nature of the test apparatus, but it appears that all three groups presented items on slides. Klinge and Rodziewicz used two rows of two pictures each, while Overton and Scott used a single row of pictures. Klinge and Rodziewicz used a four-key answer box to record answers; Knights, Richardson, and McNarry as well as Overton and Scott used touch-sensitive screens. Klinge/Rodziewicz and Overton/Scott had both the manual and the automated versions of the tests administered in the standard untimed manner; Knights, Richardson, and McNarry imposed a two-minute time limit on each item for the automated presentation only. No statistically significant main effect differences in score means from conventional versus automated presentations were found in any of these studies.[1] In each study, however, the mean scores obtained from the automated procedures were slightly lower than those from the manual administrations. The mean IQ in Knights/Richardson/McNarry and Overton/Scott was approximately 2 points lower when measured for the automated test than it was for the conventional test.

Three studies (Elwood 1972b, 1972d; Hedl, O'Neil, and Hansen 1971) used tests that contained free-response items. For free-response items, both administration and scoring procedures can affect the equivalence of scores in the two test administration modes. Because subjects are required to type in their complete free-response answers (a word, a sentence, several sentences) rather than select one of several possible keys or buttons, data entry errors and the demands of interacting with the automated equipment may result in lower scores. Once entered, responses can be saved and later scored by hand (this was done in the two Elwood studies) or scored automatically by computer (this was done in the Hedl/O'Neil/Hansen study). If the computer is used, a scoring algorithm and program must be produced, which will provide the same results as conventional scoring procedures. Differences in scoring could also produce automated scores that were not equivalent to scores from conventional procedures.

To administer the WAIS, the Elwood studies utilized a set of elaborate automated equipment that included slides, a speaker, a tape recorder, a special keyboard, and a set of drawers containing various test materials. As mentioned above, scoring of free-response items was not done by computer. In the 1972b study Elwood compared the scores on his automated test with scores on the conventional version of the WAIS and found Full Scale IQ scores to be significantly lower for the automated version. In a second study, however, there were no statistically significant differences.

The largest differences in score means between automated and conventional versions of the test (over one standard deviation) were found in the Hedl/O'Neil/Hansen study. The authors administered the Slossen Intelligence Test via an IBM 1500 Instructional System, which consists of a cathode-ray tube, a light pen, and a typewriterlike keyboard. The test was both administered and scored by computer. Thus, the large differences could be due to either the complications introduced by the automated presentation or the disagreement between conventional and computer scoring of answers to free-response questions, or both.

As stated above, only two studies in this section report results in which score means from automated versions of tests were not consistently lower than the means from conventional presentations. Curtis and Kropp (1961) administered the School and College Ability Tests (SCAT) to ninth graders. The SCAT contains both verbal and mathematical subtests, not unlike the Scholastic Aptitude Test (SAT). Items were presented on slides, which were projected to the front of the room. Answers, recorded on paper answer sheets, were hand-scored. Two versions of the automated test were given: items were displayed individually in the first version and in groups of three in the second. Performance was compared with that on one-item and three-item paper-and-pencil administrations. Curtis and Kropp report the score mean from the three-item automated verbal test was significantly higher than the mean from the corresponding conventional administration. The mean on the one-item automated verbal test was also higher than the mean on the corresponding conventional version; this difference, however, was not statistically significant. For the quantitative test no significant differences in score means were found between conventional and automated presentations. Means for the automated one-item mathematics test were slightly lower and for the three-item, slightly higher.

An additional finding of interest in the Curtis and Kropp study concerned the omitted items. For both the verbal and the quantitative tests, the mean number of omissions in both one-item and three-item automated presentations was significantly smaller than the mean number of omissions in the corresponding conventional presentations. Such differences in the omission rate as a function of test administration mode have potentially serious implications regarding score comparability for aptitude or achievement tests, which use formula scoring that involves a penalty for incorrect answers but no penalty for omitted items.

Hitti, Riffer, and Stuckless (1971) used a CRT to present the Raven's Standard Progressive Matrices to

1. Elwood and also Knights, Richardson, and McNarry did obtain a statistically significant interaction between mode of test presentation and testing order, which is described in more detail below.

deaf students. What makes this study different is that, while the directions were given and the answers were recorded by computer, the questions were presented via a paper-and-pencil booklet. Scores were higher, but not significantly so, for the automated version of this test.

Only two of the studies looked at differences between automated and conventional tests for different item types. Curtis and Kropp (1961) compared items with sentence stems and those with a stimulus word or number problem. No differences were noted for the verbal items, but an interesting interaction occurred for the quantitative items. Items that had a sentence stem produced higher score means for the automated test than for the conventional test. Items without a sentence stem, however, produced higher score means for the conventional test than for the automated test. The authors suggest this is because students tended to copy questions without a stem before attempting to answer them, as they generally consisted of a series of numbers and an operation.

Hoffman and Lundberg (1976) also found significant differences in score means between item types. These authors looked at the difference between true/false, multiple-choice, and matching items on a classroom test. Significant differences in means were found, in favor of the conventional test, on the matching items. They reasoned that this was the result of the test-taker's seeing all the matching items together on the paper-and-pencil version, presumably allowing some contextual advantage, versus seeing them one at a time on the automated version. It appears that the effects of automated testing may be specific to the item format. This issue bears close examination.

Across the 10 studies listed in Table 1, a range of correlations between automated and manual forms of the same test were obtained. The correlations for total test scores range from .67 to .96; the majority are above .80. Many of the correlations are as high as the reliability coefficients expected from similar tests and seem to indicate that the constructs being tested are similar, if not the same. In several instances, however, the correlations are surprisingly low. No particular reason could be found for the low correlation of .67 for the Coloured Progressive Matrices test given by Knights, Richardson, and McNarry. However, arguments can be made for the reasons benind some of the other low correlations reported in several of the studies.

In their 1961 study Curtis and Kropp looked at both one-item and three-item automated presentations. When putting together the triads, the authors rearranged the items according to difficulty level. Recent research has shown that item order can affect the score on a test (see Leary and Dorans 1985 for a comprehensive review). In fact, the lower correlations reported in the Curtis and Kropp study are between the paper-and-pencil version of the test and the automated version, which presented three items at a time (.86 for verbal three-item versus .93 for one-item presentation; .70 for quantitative three-item versus .79 for one-item presentation).

The relatively low correlation of .75 found in the Hedl, O'Neil, and Hansen (1971) study could be due to another reason. As discussed above, subjects taking the Slossen Intelligence Test under automated conditions were required to type in their answers. The answers were then scored by the computer, adding a variance factor of scoring reliability to the keyboard ability factor. When combined, these may be enough to reduce the correlation significantly.

An unexpected issue emerged from these early studies which is related to the design used to study the effects of automated testing. Elwood (1972b) and Knights, Richardson, and McNarry (1973) reported an interaction between test mode and test order in within-subject designs. When subjects took the automated version followed by a conventional version, greater increases in scores were found than when the conventional version was given first. Similar effects were observed in several studies reviewed in later sections. The possibility of such effects will need to be taken into account in future equating studies that use a single-group within-subject designs.

To summarize: The studies of automated testing reviewed in this section raise important concerns about the equivalence of scores from automated versions and scores from conventional versions. Despite a small number of exceptions, means were generally lower when tests were administered in an automated format than when they were administered in a conventional format. While it should be pointed out that the means for the automated tests were not always significantly lower (from the statistical standpoint), the number of subjects was generally small, with the exception of the Overton and Scott (1972) and Hoffman and Lundberg (1976) studies. As a final caveat, the practical significance of these slightly lower scores must be determined in light of the importance accorded a lower score. Whether the differences are large enough to warrant concern must be decided on an individual test basis.

Two additional issues raised by the results of the studies discussed in this section concern, first, the differences in the examinees' item omission rates in automated versus conventional presentations and, second, the effect of item type on the equivalence of automated and conventional tests. When a test is automated, free-response items may pose particular problems that are not inherent in limited-choice items. Differences between automated and conventional tests may be specific to certain types of limited-choice items.

COMPUTER-ADMINISTERED TESTS

In the current conception of an automated test, a computer controls the presentation of limited-choice items on a CRT. The studies reviewed in this section, with one or two minor exceptions, fit the current paradigm. Of the tests adapted for computer administration to date, two general groups can be distinguished. In the first group are a large number of tests consisting of limited-choice items, which can be presented either individually or in sets on a single computer screen. A second group consists of items based on extended reading passages which must be presented on a series of CRT screens. The difficulties associated with the computer administration are likely to be different for the two test groups. Therefore, studies using single-screen items and studies are reviewed separately.

Tests Consisting of Single-Screen Items

The tests that might be expected to present the least difficulty in their adaptation to a computer-administered format are those with self-contained items that can be presented on a single screen. These items do not require examinees to refer to information not presented in the item itself, such as a common reading passage or a common set of diagrams, figures, tables, and so forth. By the term *single-screen item* we mean that the question and, where applicable, all possible response options are displayed simultaneously to the examinee. For example, multiple-choice items in which the stem is displayed on one screen and response options are displayed on a subsequent screen (or set of screens) would not be classified under the heading single-screen items.

Administering single-screen items by computer is a straightforward task. The items can be displayed one at a time or in sets of two or three if they are brief. The examinee is required to indicate his or her answer by typing it on the keyboard, using a light pen, a mouse, or some other response medium. Usually a simple response such as pressing the RETURN key results in display of the next item or set of items. The examinee need not be distracted from the task of answering the question by the necessity of routing through a number of related screens to find material necessary to provide an answer.

One can further distinguish between two subcategories—tests whose problems and response options consist of only numeric or alphanumeric characters and those that contain figures, graphs, or pictures. Items containing graphics (i.e., figural or pictorial information) present certain problems not inherent in those consisting of only numeric and alphanumeric characters. For example, resolution problems associated with CRTs and TTYs may make figures more difficult to read and pictures more difficult to recognize or interpret. These factors may have an impact on performance, particularly in timed or speeded tests.

In the preceding section tests were characterized by timing conditions: (1) tests administered without a time limit; (2) tests administered with a time limit but designed to be, primarily, power tests; and (3) speed tests. One might expect, a priori, that timed power tests and speed tests would be most sensitive to mode-of-administration effects in that any distractions or difficulties introduced by computer administration, or advantages introduced by ease of responding, could affect the rate at which items are answered. Such rate differences are likely to have an effect on score comparability when tests are timed.

Untimed Tests without Graphics

Many studies have compared computer-administered and paper-and-pencil versions of untimed tests that contain items without graphics. Within this category three general classes have been studied: (1) personality inventories and psychological scales, (2) general aptitude tests, and (3) achievement tests. The results from these studies are reviewed in the following subsections.

Personality Inventories and Psychological Scales. Psychological scales and personality inventories consist of items that are usually statements describing behaviors or feelings that occur in certain circumstances. Examinees are asked to indicate whether the statements are true or false or to determine the degree to which the statements apply to themselves. During the past 15 years there has been considerable interest in adapting such standardized personality measures to administration by computer. (See Butcher 1987 for a thorough review of the literature and issues surrounding computerized psychological assessment.) In addition to the general benefits of computer delivery (for example, savings in costs and time), several authors expressed hope that computer administration might reduce examinee response biases, present in many psychological tests, such as answering in a defensive or socially desirable manner. Presumably the examinees' knowledge that their responses to personally sensitive questions are being recorded and scored by an impartial machine would decrease their fear of being evaluated negatively and increase the honesty with which they respond. Reduction in response biases might result in increased validity for such measures. However, if such effects exist, scores from computer-administered and paper-and-pencil-administered tests would most likely not be comparable and separate norming and equating studies would be required.

The nine studies reviewed in this section are sum-

marized in Table 2, which is identical in format to Table 1 with one exception: the column headed Equipment has been removed because all the studies used procedures consistent with the current conception of automated testing. All nine studies administered tests without a time limit. Six reported statistically significant differences between automated and conventional administrations. Five found differences in score means. In one study differences were observed in test-retest reliability coefficients.

Four studies—Lushene, O'Neil, and Dunn (1974), two by Biskin and Kolotkin (1977), and White, Clements, and Fowler (1985)—compared scores from computer-administered versions of the Minnesota Multiphasic Personality Inventory (MMPI) with scores from the standard paper-and-pencil mode. The MMPI consists of 566 statements to which examinees respond whether (pertaining to themselves) the statements are "true," "false," or "cannot say." Two different paper-and-pencil versions of the MMPI exist, one in which the items are administered on individual cards and a second in which the items are in booklet form. Examinee answers are used to report results on 10 basic scales: (1) hypochondriasis, (2) depression, (3) hysteria, (4) psychopathic deviation, (5) masculinity/femininity, (6) paranoia, (7) psychasthenia, (8) schizophrenia, (9) hypomania, and (10) social introversion (Anastasi 1976). In addition, 3 special scales are typically used: (1) the L scale, designed to detect the degree to which examinees are attempting to portray themselves in a favorable light; (2) the F scale, designed to detect carelessness and scoring errors in responding; and (3) the K scale, designed to detect defensiveness or an attempt to fake "good" or "bad" responses.

In the Lushene/O'Neil/Dunn study, the computer version was administered on a CRT and terminal keyboard workstation interfaced with an IBM 1500 mainframe computer. All items were presented, one at a time, on a CRT, and students responded by typing *t* (true), *f* (false), or *?* (cannot say) on a standard terminal keyboard. Pressing the space bar resulted in display of the next item.

Each of the 63 female students who voluntered to participate was randomly assigned to one of two groups. Group 1 took the computer version first and the booklet paper-and-pencil version one week later. Group 2 took the booklet paper-and-pencil version first and the computer version one week later. The paper-and-pencil test was administered to each group in one common session. It is not clear in the original article whether the computer version was administered in a group or an individual setting.

Scores were calculated for each student, in each administration mode, for each of the 10 basic and 3 special MMPI scales. Correlations between scores ranged from a low of .42 to a high of .92. Lushene/O'Neil/Dunn performed a series of discriminant function analyses on score profiles and report a number of statistically significant results. For the group that took the computer version first, statistically significant differences were found between mean scores from the computer and paper-and-pencil versions on 8 of the 13 scales. However, no consistent pattern of results emerged. For 4 of the 8 scales that showed significant differences, mean scores were higher when the test was administered by computer; for the remaining 4, mean scores were higher when the test was administered in paper-and-pencil form.

A consistent pattern of results was observed for the group that took the paper-and-pencil version first. Statistically significant differences between mean scores were found on 10 of the 13 scales, and for all of these, mean scores were higher in the paper-and-pencil than in the computer-administered test. In fact, even for the 3 scales that did not show significant differences, mean scores were higher in the paper-and-pencil mode. Seven of the 13 scales showed statistically significant mean differences for both the group that took the paper-and-pencil test first and the group that took the automated test first. However, the direction of the difference was the same for only 4 of the 7 scales.

Because of the within-subject design of the Lushene/O'Neil/Dunn study, score differences between the first and second administrations for each of the two groups are due to the joint effects of the different administration modes and the examinee experience of taking the same test twice. From the perspective of comparing computer and paper-and-pencil tests, it is helpful to examine the scaled score means obtained in each of the modes on each of the scales, ignoring, for the moment, order of administration. We've computed means for the computer administered and paper-and-pencil-administered methods, ignoring administration order, for each of the 13 scales. Means were recomputed as the unweighted average of the score means reported separately by order of administration in the original article. Thus, the recomputed means might be thought of as those that would have resulted if the subjects were equally divided between the two administration orders. The recomputed means are given in Table 3. For 11 of the 13 scales, the scores were higher when the test was administered in the paper-and-pencil mode, although for one scale the scores were almost identical. For 2 scales the scores were slightly higher when the test was administered by computer.

A set of two studies that compared a computerized version of the MMPI (similar to that used by Lushene/O'Neil/Dunn) with a standard test booklet administration was reported by Biskin and Kolotkin

Table 2. Summary of Studies Using Untimed Personality Inventories or Psychological Scales (Single-Screen Items without Graphics)

			Experimental Design				*Results*	
				Controlled for:				
Study	*Test*	*Type*	*Mode Factor*	*Order*	*Form*	*Subjects*	*Significant Differences*	*Statistics Affected*
Lushene, O'Neil, & Dunn (1974)	MMPI	Untimed	W*	Y‡	Y	—	Y	Mean
Biskin & Kolotkin (1977) Study 1	MMPI	Untimed	B†	—	—	Y	Y	Mean
Biskin & Kolotkin (1977) Study 2	MMPI	Untimed	B	—	—	Y	Y	Mean
Scissons (1976)	California Psychological Inventory	Untimed	B	—	—	Y	Y	Mean
Katz & Dalby (1981)	Fundamental Interpersonal Relations Orientation	Untimed	B	—	—	Y	Y	Test-retest reliability
White, Clements, & Fowler (1985)	MMPI	Untimed	W	Y	Y	—	Y	Mean
Koson, Kitchen, Kochen, & Stodolosky (1970)	Scales constructed from MMPI items	Untimed	B	—	—	Y	N§	—
Parks, Mead, & Johnson (1985)	Marital Adjustment Test	Untimed	W	Y	Y	—	N	—
Lukin, Dowd, Plake, & Kraft (1985)	Therapeutic Reactance Scale	Untimed	W	Y	Y	—	N	—
	State-Trait Anxiety Inventory	Untimed					N	—
	Beck Depression Inventory	Untimed					N	—

*W = within-subject variable.

†B = between-subject variable.

‡Y = yes

§N = no

Table 3. Means for Paper-and-Pencil and Computer Versions of the MMPI (Lushene, O'Neil, and Dunn 1974)

Scale	*Paper-and-Pencil*	*Computer-administered*
Hypochondriasis	5.45	5.65
Depression	20.02	19.00
Hysteria	20.54	17.82
Psychopathic deviation	15.17	13.59
Masculinity/ femininity	39.67	39.15
Paranoia	10.19	8.15
Psychasthenia	16.72	16.67
Schizophrenia	15.12	14.59
Hypomania	18.23	17.65
Social introversion	29.73	25.70
L (good self-report)	2.61	2.03
F (carelessness)	5.14	5.34
K (defensiveness)	13.57	9.44

(1977). In the first of the pair of studies, two computer-administered versions were used, one in which items and instructions were viewed on a CRT and another in which items and instructions were presented on a TTY. As in the Lushene/O'Neil/Dunn study, students in the computer conditions indicated their responses by typing *t* (true), *f* (false), or *?* (cannot say) on a standard terminal keyboard.

A total of 126 male undergraduate psychology students participated in the testing. Each subject was randomly assigned to one of three test-taking conditions—paper-and-pencil, computer-CRT, or computer-TTY. All subjects were administered 425 MMPI items used in scoring the 13 scales and the "cannot say" (or *?*) scale. Results for the cannot-say scale were not reported by Lushene/O'Neil/Dunn.

When a between-group multivariate analysis of variance (MANOVA) was conducted on the experimental data from all 14 scales, a statistically significant multivariate *F* ratio was obtained. Subsequent univariate ANOVAs were run on each of the 14 scales. Statistically significant mean differences were obtained on the *?* scale and the hysteria scale.

We calculated effect sizes[2] for each of the MMPI scales in the two Biskin and Kolotkin studies; they are reported in Table 4 . In the first experiment two aspects of the results are important: First, 10 scales have negative effect sizes (that is, the means in the computer mode are lower) in the computer-CRT condition, and 11 scales have negative effect sizes in the computer-TTY condition. As in the Lushene/O'Neil/Dunn study, there is a general trend for scores to be lower when the test is administered by computer. One glaring exception, however, is the effect size for the *?* scale, where the values are positive (that is, higher scores in the computer mode) and considerably greater in magnitude. Thus, it appears that examinees chose the cannot-say option with much greater frequency in the computer test (CRT and TTY) than they did in the paper-and-pencil test.

In attempting to explain their results, Biskin and Kolotkin noted that in the paper-and-pencil mode a cannot-say response is designated by leaving a blank space for the item. When the test is automated, a cannot-say response must be indicated by hitting the *?* key followed by the RETURN key. The researchers speculated that having to make an active response (pushing two keys) to indicate a cannot-say answer may have an effect on the frequency with which the response is chosen and may be partly responsible for the lower scores in the computer condition.

To further investigate their hypothesis, Biskin and Kolotkin conducted a second experiment. Instructions for the paper-and-pencil administration were modified so that students were directed to respond with *?* to an item they wished to omit. The computer instructions were modified so that only the RETURN key had to be pressed to indicate a cannot-say response. Biskin and Kolotkin expected that a difference would now be observed on the *?* scale, with the paper-and-pencil administration giving higher mean values than would the computer administration.

Thirty-nine male psychology undergraduates were

Table 4. Effect Sizes for Computer and Paper-and-Pencil Forms of the MMPI (Biskin and Kolotkin 1977)

Scale	*Effect Size Study 1/CRT*	*Effect Size Study 1/TTY*	*Effect Size Study 2*
?	4.62	4.32	−.08
L	.05	.15	−.22
F	.08	−.20	.15
K	−.20	.16	.11
Hypochondriasis	−.09	−.19	−.27
Depression	−.18	−.14	−.20
Hysteria	−.49	−.01	−.44
Psychopathic deviation	−.36	−.20	−.14
Masculinity/ Femininity	−.33	−.34	−.38
Paranoia	−.48	−.32	−.60
Psychasthenia	−.15	−.19	−.17
Schizophrenia	−.08	−.22	−.33
Hypomania	−.19	−.08	−.08
Social introversion	.24	−.14	.16

2. Throughout the paper, *effect size* is defined as the mean score obtained by the computer-administered group minus the mean score obtained by the paper-and-pencil-administered group divided by the standard deviation of scores obtained by the paper-and-pencil-administered group.

randomly assigned to either the modified paper-and-pencil or the modified computer-CRT version. As in the first experiment, each subject was administered the 425 items needed to score the scales. When a between-group MANOVA was conducted on the data from the second experiment, the *F* ratio did not indicate a statistically significant difference between the group mean vectors.[3] More important, despite expectations, the average number of cannot-say responses in both conditions were quite similar to the computer-CRT and computer-TTY values in the first experiment. Biskin and Kolotkin conjectured that the activity in making a RETURN response in the computer-CRT test is comparable with that of indicating a *?* in the paper-and-pencil test. Further, Biskin and Kolotkin interpreted the lack of statistical significance for the miltivariate *F* test as support for their hypothesis that any scale differences were due to differences in the rates of cannot-say responses in the two versions of the test.

The effect sizes we obtained for each of the MMPI scales based on the data from the second experiment are also displayed in Table 4. It is important to note that despite similar rates of cannot-say responses, negative effect sizes were still obtained for 10 of the 13 scales studied by Lushene/O'Neil/Dunn. Thus, once again, taken as a whole, results suggest that MMPI scale scores may be slightly lower when the test is administered by computer.

A third group of researchers compared scores from a computer-administered MMPI with scores from a paper-and-pencil administration. This study was reported by White, Clements, and Fowler (1985), who created two different computer versions of the MMPI. Both were administered by Apple II microcomputers with standard keyboards and monitors. One of the computer versions (CV1) was identical to that used in their first experiment by Biskin and Kolotkin (1977). In the second computer version (CV2), the cannot-say response option was removed; subjects could respond with true or false only.

The subjects were 150 undergraduates (83 females and 67 males). They volunteered to participate in the experiment, which required them to take the MMPI on two separate occasions, and were randomly assigned to one of six groups of 25 subjects each. To the extent that it was possible, the percentage of males was kept constant across groups. The first group was tested twice with the paper-and-pencil version (booklet format). The second and third groups were tested once with the paper-and-pencil version and once with CV1 with the order of administration counterbalanced across the two groups. The fourth and fifth groups were tested once with the paper-and-pencil version and once with the CV2 with the order of administration counterbalanced across the two groups. The sixth group was tested twice with CV1. The time between testings ranged from 6 to 20 days.

Unlike the scores in the previous studies of the MMPI, the scores in this study were not consistently lower in the computer administrations; only about half the scales were lower. Statistically significant effects for mode of administration were found on only two scales, one of which was the hypochondriasis scale. The second was the *?*, or cannot-say, scale, on which effects were found in the Biskin and Kolotkin studies. As in the Biskin and Kolotkin studies, the mean number of *?* responses was dramatically higher in the CV1 than in the paper-and-pencil format. The score means on the *?* scale for the two groups who were administered the test by CV1 at the first testing session were 15.96 and 14.36. The corresponding means for the three groups administered the test in the paper-and-pencil format at the first testing session were 2.04, 0.96, and 0.44. (Score means for subjects in the CV2 condition were constrained to be 0 in the CV2 condition.) The score means on the *?* scale for the two groups who were administered the test by CV1 at the second testing session were 10.24 and 7.48. The corresponding means for the three groups administered the test in the paper-and-pencil format at the second testing session were 1.24, 1.56, and 0.00.

Like the results previously reported by Curtis and Kropp (1961), the results both of the two Biskin and Kolotkin studies and of the White, Clements, and Fowler study have some particularly important implications for the administration of formula-scored achievement tests. If the phenomena Curtis and Kropp noted with the School and College Ability Tests and the MMPI are not peculiar to their experiments, and their explanations are correct, the frequency with which examinees omit items may be influenced by the mode of delivery. For formula-scored tests, where examinees are penalized for incorrect answers but not for omitted items, different omit rates are likely to result in scores from computer and paper-and-pencil versions that are not comparable. It should be noted that in the Curtis and Kropp study, omit rates in an automated presentation were lower than those in a conventional administration. These results are in the opposite direction from those obtained by Biskin and Kolotkin and by White, Clements, and Fowler. Research will probably be needed to determine the nature of the differing results and the best method of implementing automated tests to minimize these effects.

Another self-report inventory, which contains items similar in format to the MMPI, is the California

3. Subsequent univariate tests did, however, produce one statistically significant contrast.

Table 5. Effect Sizes for Computer and Paper-and-Pencil Forms of the CPI (Scissons 1976)

Scale	*Effect Size*
CS	−.58
Well-being	−.85
Responsibility	−.87
Self-control	−.88
Tolerance	−.70
Good impression	−.99
Achievement via conformity	−.65
Achievement via independence	−.78
Intellectual efficiency	−.70
Py	−.55
Feminine	−.56

Psychological Inventory (CPI). Scissons (1976) conducted a study comparing a computer administration of the CPI with a standard paper-and-pencil administration. The former was administered at IBM 1510 student consoles interfaced to an IBM 1500 mainframe computer system. Instructions and items were presented on the consoles, and students responded (yes, no, or omit) with a light pen. The computer system gave the students practice in using the device and alerted them when erroneous light pen responses were made.

Participants were 40 male and 40 female students, randomly selected from a pool of over 200 introductory educational psychology students who volunteered for the study. The male and female groups were each randomly divided into two groups of 20. One group of males and one group of females took the computer-administered CPI; the two other groups took the paper-and-pencil version. Two male students and two female students in the paper-and-pencil groups failed to complete the CPI and were removed from the study.

Scores were determined for each student on all 18 scales of the CPI. For the male groups, statistically significant mean differences were found for 8 of the 18 scales. In each case, score means on the computer version were lower than those on the paper-and-pencil version. For the female groups, a statistically significant difference was found only for the femininity scale, and again, the mean in the computer version was lower than that in the paper-and-pencil test.

Scissons reports means and standard deviations only for males and only for 11 of the 18 scales. As a consequence, we could not determine effect sizes for all CPI scales. Effect sizes were determined for those scales for which results were reported, however, and are displayed in Table 5. The CPI results reported by Scissons, obtained with a similar psychological inventory, are consistent with those in the MMPI studies. All effect sizes are negative, indicating a general trend for scores to be slightly lower when the test is administered by computer than when it is administered in the standard paper-and-pencil fashion.

To summarize: Despite differences in experimental design—mode of test administration was a within-subject effect in Lushene, O'Neil, and Dunn and a between-subject effect in Biskin and Kolotkin and in Scissons—these studies, taken as a set, seem to provide evidence that for a personality inventory like the MMPI or the CPI, a computer presentation may result in somewhat lower scores. If so, separate norms for score interpretations based on computer-administered inventories may be required. It should be kept in mind, however, that many, if not most, of the mean differences in the Lushene/O'Neil/Dunn, Biskin/Kolotkin, and Scissons studies are not statistically significant and may be small enough to make little or no practical difference.

A number of studies comparing automated and conventional versions of several other personality measures do *not* show effects for mode of presentation, at least with respect to the mean scores. One such study, in which the effects of computer administration of personality inventory items were investigated, is reported by Koson et al. (1970). They created a personality inventory questionnaire containing three sets of questions. The first set, which Koson et al. refer to as "threat" questions, was selected from the MMPI and had been rated by 20 subjects as most threatening or personally embarrassing. These items were keyed so that higher scores indicated more honest answers; Koson et al. refer to this set of items as the "honesty" scale. A second set of items selected from the MMPI was rated by the same subjects as "neutral." The third set of questions was selected from the MMPI K scale, described previously as having been designed to detect defensiveness. In the study, K-scale items were keyed so that high scores indicated defensive responding.

Forty-eight University of Michigan students were assigned to three groups of 16. For one group, questions were presented individually on a CRT interfaced with a DEC PDP-8 minicomputer. Examinees indicated responses by depressing appropriately marked levers; no experimenter was present during the administration. In the second group, the examinee filled out a paper-and-pencil questionnaire in the presence of an experimenter. In the third group, examinees responded verbally to the questions, which were read by an experimenter.

Score means were obtained for each of the three groups on the honesty and defensiveness scales. In addition, the total number of neutral statements each group agreed with was determined. No statistically significant differences were obtained on any of the scales. On the average, examinees in the computer group had slightly higher (more honest) scores than did the paper-and-pencil group. The effect size for the mean difference was 0.32, and the standard deviation of scores was approximately 33 percent greater for the group who were

administered the questionnaire by computer. A similar pattern of results was obtained for the K-scale questions. The computer group displayed less defensive responding than did the paper-and-pencil group. However, the magnitude of the effect was extremely small, 0.07. As with the honesty scores, the standard deviation for the computer group was about 30 percent greater than that obtained by the paper-and-pencil group. All three groups responded about the same on the neutral questions.

Katz and Dalby (1981) compared the performance of 40 gifted children and 40 behavioral-problem children on the Fundamental Interpersonal Relations Orientation (FIRO) personality questionnaires. The FIRO yields scores on each of six scales. Half the gifted children and half the behavioral-problem children were administered the questionnaires via a PDP 11/70 time-sharing computer system. The remaining children were administered the questionnaires in the standard paper-and-pencil format. Both groups were retested, in the same test administration mode, one week later. There were no significant differences in score means on any of the scales. However, for three of the six scales significantly higher test-retest reliability coefficients were obtained from the computerized presentation than from the paper-and-pencil version. The paper-and-pencil mode had a significantly higher reliability for one of the scales.

A computerized version of the Marital Adjustment Test (MAT) was prepared by Parks, Mead, and Johnson (1985). The paper-and-pencil test is a 15-item instrument designed to measure an individual's perceived adjustment to his or her marriage. Unfortunately, no details concerning the items were given by the authors. In the automated test the items were presented on an Osborne Executive portable computer with a 7-inch video monitor and a typewriterlike console. Fifty couples were randomly selected from a large university's married-student housing unit. Each couple was then randomly assigned to one of two groups. Group 1 took the paper-and-pencil version first, followed by the computer-administered version. Group 2 took the tests in the reverse order. All testing was done in the couple's home, and each member of the couple responded independently to the questionnaire.

For each of the experimental groups, score means on both versions were almost identical, and the differences were also not statistically significant. The correlation between scores on the two tests was 0.98 for the combined group, leading Parks, Mead, and Johnson to conclude that scores from the two versions could be used interchangeably. A potential problem, however, with the study is that both tests appear to have been administered consecutively in a single session. With a test of only 15 items, it would seem likely that examinees would remember their responses from the initial administration and that such a carryover effect might make scores in the two modes appear more comparable than they might under more independent retest conditions.

Lukin et al. (1985) compared computer-administered versions of three psychological assessment instruments with their paper-and-pencil counterparts. The first was the Therapeutic Reactance Scale (TRS), a 32-item scale developed to assess client compliance with therapist directives. The second was the State-Trait Anxiety Inventory (STAI), which consists of two 20-item scales, one intended to measure situation-specific anxiety and the other intended to measure long-term anxiety. The third was the Beck Depression Inventory (BDI), a 21-item depression scale.

The automated forms of each of the three tests were administered on a TRS-80 model III microcomputer. Items were presented one at a time, and examinees indicated their responses by depressing a numeric key on the keyboard, followed by the RETURN key. Subjects were given an oppportunity to change their responses. (The authors do not indicate whether they could do so only for the item just presented or whether they could review and change answers for previous items.)

Sixty-six undergraduates (44 female and 22 male) from an introductory educational psychology course took both the paper-and-pencil and the computer versions of each of the three instruments. Half the students were given the paper-and-pencil version of all three tests in one session and were then given the computer version of each test one week later. The other half received the computer version of each test first, followed one week later by the three paper-and-pencil tests. Within a testing session, the order of test administration was counterbalanced following a Latin square design.

The score means on each of the instruments were almost identical for both automated and conventional administrations. The largest difference was only 0.29 points on an 80-point scale. A 2-by-2 Latin square ANOVA indicated that none of these differences was statistically significant.

In sum, the results from research on the computer administration of untimed personality tests using single-screen items are mixed. Clearly, instances of comparable scores from computer and paper-and-pencil versions of the tests can be found. However, the research on computer modes of the MMPI and the CPI suggests that score differences may be due to the mode of administration. When computer versions of standardized tests are created, careful attention has to be paid to the technique by which examinees indicate their preference to omit an item.

General Aptitude Tests. A second class of untimed tests that use single-screen items are what might be termed

Table 6. Summary of Studies Using Untimed Aptitude Tests (Single-Screen Items without Graphics)

			Experimental Design				*Results*	
				Controlled for:				
Study	*Test*	*Type*	*Mode Factor*	*Order*	*Form*	*Subjects*	*Significant Differences*	*Statistics Affected*
Johnson & Mihal (1973)	School and College Ability Tests	Untimed	W*	Y‡	Y	—	Y	Mean
Lee & Hopkins (1985)	ASVAB arithmetical reasoning test	Untimed	W	N§	N	—	Y	Mean
Lee, Moreno, & Sympson (1986)	ASVAB arithmetical reasoning test	Untimed	B†	—	—	Y	Y	Mean

*W = within-subject variable.

†B = between-subject variable.

‡Y = yes

§N = no

tests of general verbal or mathematical aptitude. These tests often consist of multiple-choice items such as vocabulary problems, analogy problems, or simple arithmetic problems. They differ from achievement tests in that the items are not designed to test knowledge or concepts clearly tied to an identified curriculum.

Three studies were identified in which automated forms of untimed general aptitude tests were compared with their paper-and-pencil counterparts. The results, summarized in Table 6, all show statistically significant differences in score means.

In the first study Johnson and Mihal (1973) compared the performance of black and white examinees on a paper-and-pencil form of the School and College Ability Tests, developed by Educational Testing Service (ETS), with an alternate form administered by computer. The SCAT contains a 50-item verbal subtest and a 50-item quantitative subtest. In the computer administration, examinees responded on dial-up typewriter terminals interfaced to an IBM 360 time-sharing system. No details concerning the rate at which the items were presented, or whether the examinees could omit or change answers, are given in the article describing the study.

Male seventh- and eighth-grade students (10 black and 10 white) were tested in pairs. Within each pair one student was administered the computer version first, while the second student received the paper-and-pencil version. Upon completing their respective tests, each student took the test in the alternate mode.

Computer administration resulted in significantly higher total SCAT scores for blacks but not for whites. The average total test score of the black students was 5.2 points higher (on a 100-item test) on the computer form than on the paper-and-pencil form. The analogous difference for the white students was about 0.5 points. Johnson and Mihal also analyzed their results separately for the verbal and quantitative subtests. The black students, on the average, scored between 3 and 4 points higher on the computer-administered verbal subtest than they did on the paper-and-pencil subtest, and this difference was statistically significant. The white students, on the average, scored between 2 and 3 points lower on the computer-administered verbal subtest, but this difference was not statistically significant. Both blacks and whites, on the average, scored between 1 and 2 points higher on the computer form of the quantitative subtest, and differences were statistically significant. Johnson and Mihal interpret their results as evidence of an increased motivation on the part of black subjects because of the novelty of the computerized testing environment as well as a reduction of the negative expectations sometimes associated with intelligence and aptitude measurement.

In a study by Lee, Moreno, and Sympson (1986), the performance of 585 naval recruits on a computerized version of an experimental arithmetical reasoning test (E-ART) was compared with their performance on a paper-and-pencil version of the same test. The E-ART was created from items like those used in the Armed Services Vocational Aptitude Battery (ASVAB) arithmetical reasoning test. All recruits had taken an operational paper-and-pencil version of the ASVAB-ART two to six months before the E-ART study. In the study each recruit was randomly assigned to either the paper-and-pencil administration or the computer administration. Subjects using paper-and-pencil were tested in groups of four to ten. Each received a test booklet containing 30 items with approximately 8 items on a page. Subjects using a computer were tested in groups of four at separate terminals in one large room; questions were presented one at a time, and subjects could not refer to previous items or change answers. In addition, subjects were not permitted to omit items when taking the E-ART in either group. Although the ASVAB-ART is usually administered as a timed test, no time limit was imposed on the E-ART subjects in either group. The authors of the study hypothesized that under totally

unspeeded conditions, no differences would be obtained between the two administration modes.

The mean and standard deviation of number-right scores on the ASVAB-ART were identical to one decimal place for the two groups in the study. In contrast, the group who took the computer version of the E-ART scored about 1 point lower (on a 30-item test) than the paper-and-pencil group. The difference in means is statistically significant. In addition, the proportion answering correctly in the paper-and-pencil group was larger for 21 of the 30 items than the proportion of recruits who received the test by computer. Only 3 items displayed difficulty effects in the opposite direction.

A possible reason that computer scores were lower than paper-and-pencil scores may be that examinees did not have an opportunity to review previous answers and to correct erroneous responses. If a computer examinee made a key entry error (for example, wanted to answer response option *b*, but inadvertently depressed response option *c*), no opportunity was provided to correct the mistake. Such opportunities clearly did exist for paper-and-pencil examinees and may be partly responsible for the statistically significant result.

Lee and Hopkins (1985) report a study comparing paper-and-pencil and computerized versions of an arithmetical reasoning test almost identical to the test used in the preceding discussion. The results were similar to those reported by Lee, Moreno, and Sympson. Two 30-item forms were developed, one administered in a paper-and-pencil mode and the second in a computerized version. The computer test was given on an APPLE II+ microcomputer. The questions were displayed individually on a CRT, and the subjects typed their answers on a keyboard. All subjects took the paper-and-pencil version first and were tested individually in both conditions. No time limits were imposed. In addition, as in the preceding study, subjects in either arrangement were not permitted to omit items. Also, as in the preceding study, subjects using a computer were not permitted to review items or change answers.

As in the Lee/Moreno/Sympson study, score means were lower on the computerized test. The mean number-right score in the paper-and-pencil version was almost 2 points higher (on a 30-item test) than that achieved in the computer administration, and this difference is statistically significant. In the discussion section of their article, Lee and Hopkins remark that some examinees indicated they made key errors in entering responses; these may have been responsible, to some degree, for the observed mean difference. Unlike the previous study, however, the effect of test mode administration is perfectly confounded with test form. The different test forms used were created by the authors for the study. They were not previously created alternate forms that had been demonstrated to be parallel in difficulty. In addition, because the computer versions were always administered second, whatever effects are associated with repeated testing are also confounded with mode-of-administration effects. Thus, the results of the Lee and Hopkins study need to be interpreted cautiously.

In sum, despite relatively simple item types, mode-of-administration effects were found in all three studies of untimed aptitude tests with single-screen items. In one study means tended to be higher on the computer-administered version than on the paper-and-pencil version, perhaps because of differences in motivation or interpersonal considerations associated with paper-and-pencil ability measures. In two of the three studies, scores on the computer-administered version of the test were lower than scores on the corresponding paper-and-pencil version. The lower scores may be due, at least partly, to the constraint imposed in the computer mode on reviewing and changing answers. For one of the studies, however, it might be equally justified to attribute the lower scores in the computer administration to differences in the difficulty of the forms used or to the effects of repeated testing.

Achievement Tests. Three studies compared automated and conventional achievement tests administered without a time limit. The results are summarized in Table 7. None of the three presents evidence of mode-of-administration effects.

As part of a larger study that included computer-adaptive tests, Olsen et al. (1986) compared computer-administered versions of the California Assessment Program mathematics applications tests with paper-and-pencil versions. Third-grade and sixth-grade students from several California school districts were randomly assigned to a number of different groups. Each group was administered an appropriate grade-level test on two occasions. The third-grade test contained 55 items, and the sixth-grade test contained 62 items.

Each group took the test once in an IRT computerized adaptive testing mode and a second time either with standard paper and pencil or by computer. In the computer-administered form, the entire set of items in the paper-and-pencil version was presented sequentially on the CRT of a PLATO/WICAT System 300 microcomputer system. The students entered their responses on the computer keyboard. No details are provided, however, on whether they could review items or correct their answers. In the IRT computer-adaptive condition, between 15 and 20 items were administered in an adaptive mode. At each grade level one group of students took the adaptive test first, followed by the computer-administered test. A second group of students took the computer-administered test, followed by the adaptive test. A third group of students took the computer-

Table 7. Summary of Studies Using Untimed Achievement Tests (Single-Screen Items without Graphics)

			Experimental Design				Results	
				Controlled for:				
Study	*Test*	*Type*	*Mode Factor*	*Order*	*Form*	*Subjects*	*Significant Differences*	*Statistics Affected*
Olsen, Maynes, Slawson, & Ho (1986)	California Assessment Program mathematics applications	Untimed	B*	—	—	Y†	N‡	—
Wise & Wise (1986)	Classroom arithmetic test	Untimed	B	—	—	Y	N	—
Wise, Boettcher, Harvey, & Plake (1987)	Classroom algebra test	Untimed	B	—	—	Y	N	—

*B = between-subject variable.

†Y = yes

‡N = no

adaptive test followed by the paper-and-pencil version, while a fourth group took the tests in the reverse order.

A total of 263 third graders and 171 sixth graders took the computer-administered versions, while different groups of 214 third graders and 129 sixth graders took the paper-and-pencil versions. Analyses were conducted separately for the third-grade and sixth-grade data. No significant differences were found between score means obtained on the paper-and-pencil and computer versions for either the third- or the sixth-grade groups. In fact, means were very similar for the third graders (36.57 versus 35.76 for the computer and paper-and-pencil administrations, respectively) and almost identical for the sixth graders (30.47 versus 30.49). In addition, Olsen et al. computed coefficient alpha reliabilites and average standard errors of measurements for the two versions at each grade level. The results of the paper-and-pencil and the computer administrations were almost identical at both grade levels, differing (if at all) only in the second decimal place.

Wise and Wise (1986) compared the performance of 68 third- and fourth-grade students on paper-and-pencil-administered and two computer-administered versions of a 32-item multiple-choice test of basic arithmetic skills. The computer versions were administered on the APPLE II+ and APPLE IIe microcomputers. In one computer version, immediate feedback on the correctness of a response was provided. In the second computer version—the one of interest from the point of view of our review—no feedback was provided. The subjects, all of whom had substantial experience using microcomputers at school, were randomly assigned to one of three administration procedures: paper-and-pencil, computer with no feedback, and computer with feedback.

In addition to the experimental test scores, each subject's scaled score from a prior administration of the mathematics computation subtest of the California Achievement Tests was obtained. These scores were used to form two groups, high mathematics achievement and low mathematics achievement, based on a median split. Number-right arithmetic skills scores were analyzed in a three-modes-of-administration by two-levels-of-achievement ANOVA. No significant main effects or interactions were obtained. For both the high- and low-achievement test groups, the score means of those students administered the test in paper-and-pencil mode was within 0.3 points (on a 32-item test) of the mean of those students in the no-feedback computer mode. The standard deviations of scores, however, were about 82 percent greater for the groups taking the test in paper-and-pencil format than they were for the group taking the test in the no-feedback computer condition. This result should be viewed with caution because the sample sizes for each group are extremely small (fewer than 15 in a group).

Wise et al. (1987) conducted a study to investigate the effects of two possible individual difference variables, computer anxiety and computer experience, on the comparability of scores from paper-and-pencil and computer versions of a basic algebra skills test. The algebra test consisted of 25 four-option multiple-choice items. Two forms of the test were used—one with items ordered from easy to hard and the second with items ordered randomly. Two versions of each form were created—a paper-and-pencil version and a computer version. The computer test was administered by an APPLE IIe microcomputer.

A total of 100 students (24 males and 76 females) from an introductory college statistics course participated in the study. Each student was randomly assigned to one of four groups: ordered/paper-and-pencil, random/paper-and-pencil, ordered/computer,

random/computer. The subjects were tested in groups of four in both computer and paper-and-pencil versions of each test form. All students completed a computer anxiety scale and estimated the total number of hours they had spent using computers prior to taking the algebra skills test. They completed the State-Trait Anxiety Inventory (STAI) after taking the algebra skills test. The students were assigned to one of two anxiety groupings (low versus high), based on a median split on scores from the computer anxiety scale, and to one of two experience groupings (no experience versus experience), based on the number of hours of computer experience they reported.

Data were analyzed in a two-item-arrangements by two-response-modes by two-anxiety-levels MANOVA with number-right scores on the algebra test and STAI scores serving as the two dependent variables. A second MANOVA was performed with computer experience instead of computer anxiety as the third independent variable. No statistically significant effects were obtained. For the easy-to-hard item arrangement, the score means were slightly higher for the group that received the tests in a paper-and-pencil format, while the reverse was true for the test forms with the items in random order.

All three studies that used untimed achievement tests found no differences in mean test scores as a function of mode of test administration. This is surprising because differences were reported for all three studies that used untimed general aptitude tests. The differences may be related to the examinees' opportunities to correct errors. However, few details concerning answer-changing options are provided in the last three studies in this section.

Timed Power Tests without Graphics

Four studies report results in which scores from computer-administered versions of timed power tests were compared with their paper-and-pencil counterparts (Table 8). Three of the four studies report statistically significant differences in mean scores obtained under automated versus conventional conditions. The experimental design employed by one of these studies, however, makes interpretation of the results somewhat problematic.

The most relevant study—from the point of view of our literature review—is reported by Sachar and Fletcher (1978). In study 1 and study 2, two separate experiments, they compared the statistical characteristics of two types of aptitude tests when administered by paper and pencil and when administered by computer. One test consisted of two forms of a 50-item, five-option multiple-choice vocabulary test with a 15-minute time limit. The second test consisted of two forms of a 30-item, true/false, symbolic reasoning test with a 5-minute time limit. The computer versions were presented on a PLATO IV computer system with a touch-sensitive plasma panel. Items were shown one at a time, and examinees responded by touching the desired square on the panel. Examinees could skip items, review previous items, and change answers. To do so, however, they were required to go back (or forward) sequentially through each previous item to review or

Table 8. Summary of Studies Using Timed Power Tests (Single-Screen Items without Graphics)

		Experimental Design				*Results*	
			Controlled for:				
Study	*Test*	*Mode Factor*	*Order*	*Form*	*Subjects*	*Significant Differences*	*Statistics Affected*
Sachar & Fletcher (1978) Study 1	Vocabulary	W*	Y‡	Y	—	N	—
	Symbolic reasoning					Y	Mean
Sachar & Fletcher (1978) Study 2	Vocabulary	B†	—	—	Y	N	—
	Symbolic reasoning					Y	Mean
Wildgrube (1982)	Word analogies	W	N§	Y	—	Y	Mean
	Spelling					Y	Mean
	Arithmetical reasoning					N	—
McBride & Weiss (1974)	Word knowledge	B	—	—	Y	N	—

*W = within-subject variable.

†B = between-subject variable.

‡Y = yes

§N = no

change answers to items not adjacent to the one they were currently working on.

A total of 118 recruits at a Navy Training Center participated in the first experiment. Each recruit was administered one form of each of the aptitude tests in the standard paper-and-pencil format and was administered the alternate form by computer. Half the recruits received one form of both tests in the computer mode first and subsequently received the alternate form of each test in the paper-and-pencil mode. The other subjects received tests in the opposite order. The design was also counterbalanced so that each form of both tests was given in a paper-and-pencil format and a computer format an equal number of times.

Sachar and Fletcher obtained somewhat mixed results for the first experiment. For one form of the vocabulary test, the score means of the group that received the paper-and-pencil administration was about 3 points higher than the mean of the group that received the computer administration. The reverse was true for the second vocabulary test form. The computer group had a score mean about 1.5 points higher than that of the paper-and-pencil group. As a consequence, the overall difference in means in the computer and paper-and-pencil versions was not statistically significant. For the symbolic reasoning test, however, the group that was administered the computer version scored about 2 points lower than did the group that was administered the paper-and-pencil version for both forms of the test. Differences between means for the symbolic reasoning test were statistically significant.

In the second experiment, 91 recruits participated. Each recruit again took both forms of each test. However, in the second experiment approximately half the subjects took all test forms in the standard paper-and-pencil format, while the remainder took all test forms in the computer format. The vocabulary test scores for both forms were almost identical for the computer groups and the paper-and-pencil groups. The results of the symbolic reasoning test, however, paralleled those of the first experiment. The score mean was about 2.5 points higher for the group receiving the forms in the paper-and-pencil version than it was for the group receiving those forms in the computer version.

Taken as a set, the results of the two experiments show clear mode-of-administration effects for the symbolic reasoning test but not for the vocabulary test. One distinct difference between the tests, noted by Sachar and Fletcher, is in the degree of speededness. The vocabulary test requires the examinee to answer 50 questions in 15 minutes (3.33 questions per minute). In contrast, the symbolic reasoning test requires the examinee to answer 30 questions in 5 minutes (6.00 questions per minute). The demands of taking the test by computer, in particular the operations involved in reviewing previous items and changing answers, may sufficiently slow performance, particularly in the highly speeded symbolic reasoning test, to result in decreased scores on the computer version.

Wildgrube (1982) reports the results of an initial pilot study comparing paper-and-pencil versions with microcomputer versions of six subtests from the German Federal Armed Forces Aptitude Classification Battery (EVT). The subtests were a word analogy test, a figure reasoning test, an arithmetical reasoning test, a spelling test, a mechanical aptitude test, and an electrotechnical comprehension test. Three of the tests, word analogies, arithmetical reasoning, and spelling, can be classified as timed power tests containing no figural or pictorial material.

Each microcomputer testing station had a screen, a microprocessor, and a special keyboard containing the keys necessary for examinees to respond. Few details were given on the specifics of how the testing system operated. A sample of 208 army recruits who were tested with paper-and-pencil versions of the EVT were retested (several weeks to several months later) with parallel forms of the computer-administered subtests. All tests had the same time limits as their paper-and-pencil counterparts. The word analogy test contained 20 items to be completed in 4.5 minutes; the spelling test contained 50 items to be completed in 3 minutes; and the arithmetical reasoning test contained 20 items to be completed in 14 minutes.

For two of the tests, word analogies and spelling, score means were higher in the paper-and-pencil versions, and these differences were statistically significant. It should be noted that based on the time limits, these two tests may be somewhat speeded. However, the Wildgrube results must be interpreted cautiously. All subjects took the computer versions several weeks after taking the paper-and-pencil versions. As a result, order effects may be confounded with mode-of-administration effects.

In an effort to develop a word-knowledge item pool to be used in computerized adaptive testing, McBride and Weiss (1974) conducted an item norming study to determine whether item parameters obtained in a pencil-and-paper administration would differ from those obtained in a computer administration. A set of 240 multiple-choice vocabulary items was administerd in one of three modes. In the first and second modes, each item was typed on a blank computer card, and the cards were arranged in individual decks of 240 items each. For each of 50 of the subjects who participated in the experiment, the item order in the decks was randomly determined. For a second group of 67 subjects, a single random order was determined. No time limit was imposed in either of the first two modes. In the third mode, items were administered individually on a CRT. Each of 75 subjects completed a random-order 240-item test. Subjects indicated their responses by pressing

Table 9. Summary of Studies Using Speed Tests (Single-Screen Items without Graphics)

Study	Test	Experimental Design				Results	
		Mode Factor	Controlled for: Order	Form	Subjects	Significant Differences	Statistics Affected
Greaud & Green (1986)	ASVAB numerical operations	W*	Y†	Y	—	Y	Mean/alternate forms reliability
	ASVAB clerical speed					Y	Mean
Kiely, Zara, & Weiss (1986)	ASVAB numerical operations	W	Y	Y	—	Y	Mean
	ASVAB clerical speed					Y	Mean

*W = within-subject variable.

†Y = yes

a light pen against the position on the screen at which the chosen alternative appeared. A 50-minute time limit was imposed.

In an attempt to remove the effects of differential speededness, the main dependent variable used by McBride and Weiss was what they termed an *accuracy score*. Accuracy scores were the ratio of correct answers to the number of items presented. The mean accuracy score did not differ significantly across the three modes of administration; however, score means were slightly higher when the items were presented on cards. The accuracy means were 0.71 and .072 for the fixed and random-order paper-and-pencil formats, respectively. The accuracy score mean for the computer format was 0.69.

In sum, the studies by Sachar and Fletcher showed clear mode-of-administration effects for the highly speeded symbolic reasoning test. A possible cause of the mode-of-administration effects is differential speededness associated with the error correction and review features of the computer version. The study by Wildgrube showed possible mode-of-administration effects; however, differences may be confounded with order effects. The study by McBride and Weiss is difficult to interpret because, for the purposes for which their study was intended, they used a dependent measure designed to remove the effects associated with differential speededness. Differential speededness is likely to be one of the major reasons that mode-of-administration effects are obtained with timed power tests.

Speed Tests without Graphics

Given the results reported in the preceding section, one would expect speed tests to be particularly sensitive to mode-of-administration effects. Such effects could be facilitative, if the manual operations associated with recording answers were less cumbersome in computer-administered versions of the test. However, negative mode-of-administration effects may be observed if the computer version is more awkward to work on. Both the method of indicating responses and the number of items presented at a time may differ in computer-administered and paper-and-pencil-administered tests. These differences will likely affect the rate of responding and thus may affect the comparability of automated and conventional scores.

Two studies were identified that compared performance on computer-administered speed tests with corresponding paper-and-pencil tests. Both studies, summarized in Table 9, report statistically significant differences in the score means obtained in the two test modes.

Greaud and Green (1986) compared performance on paper-and-pencil and computerized versions of two clerical speed tests from the ASVAB, the numerical operations test (NO) and the clerical speed test (CS). The NO test consists of simple four-option multiple-choice arithmetic problems. The CS test consists of a single master list of words, with each word adjacent to a four-digit numeric code. Below the master list is a second list of five-option multiple-choice items. The stem of each item contains one of the words in the master list. The answer options consist of four-digit numeric codes. An examinee's task is to choose the answer option that gives the four-digit numeric code associated with the stem word in the original master list.

For each test, four shortened experimental forms were created from the items in retired operational ASVAB forms. The experimental NO forms contained 25 problems; the experimental CS forms consisted of a single master list and 28 multiple-choice four-digit coding problems. For the paper-and-pencil versions, problems were presented on a single sheet of paper, and examinees coded their responses on a separate "bubble" answer sheet. For the computer version, the problems

were presented individually on a TVI-920 computer terminal, and examinees indicated their responses by pressing one of two adjacent labeled keys in the top row of the keyboard.

In both conditions, no time limits were imposed. However, the total amount of time required to complete all items and the number of correct responses were recorded for both modes. The number of correct responses divided by the total time to complete the test (CPM) was used as the dependent variable for the comparisons of the two versions of the tests. It should be noted that the administration procedure for both the experimental paper-and-pencil and computer tests in this study differs from the ASVAB operational paper-and-pencil procedure. In the latter, subjects attempt to respond to as many items as they can in a fixed interval of time.

Fifty college students took two forms of each test by paper-and-pencil and two forms of each test by computer for a total of eight tests per student. Half the group took the four paper-and-pencil forms first, and the other half took the four computer forms first.

A strong significant effect for mode of test administration on CPM scores was observed for both tests. For the NO test, the score mean was roughly six items (per minute) higher in the computer-administered mode than in the paper-and-pencil version. For the CS test, the mean score for the computer-administered version was over five and one-half items (per minute) higher than the mean in the paper-and-pencil version. Reliability coefficients for the NO test were significantly higher in the computer condition (.77) than they were in the paper-and-pencil version (.62). No significant differences in reliability coefficients were observed for the two versions of the CS tests (.70 for paper-and-pencil versus .74 for computer). In addition, Greaud and Green estimated the "true score" correlations between computer and paper-and-pencil versions. These correlations ranged from .81 to .98, which, although high, are far enough below 1.0 in at least one instance to suggest that slightly different constructs are being measured in the two different test modes.

A second study of computer-administered versions of the ASVAB tests was conducted by Kiely, Zara, and Weiss (1986). Their study compared paper-and-pencil versions of six ASVAB tests with several different computer-administered versions. Two of these six tests were the clerical speed measures (NO and CS) studied by Greaud and Green. These tests, in their standard administration format, require subjects to respond to as many items as they can in a fixed time.

Two different computer versions of the NO and CS tests were prepared. In the first, CRT-1, items appeared on the screen one at a time as in the Greaud and Green study. In the second, CRT-2, several items appeared on the screen at the same time. For the NO test there were three items at one time; for the CS test, two blocks of seven items were displayed.

For the testing sample, 997 Air Force recruits were divided into three general experimental groups. Each recruit took all six ASVAB tests in alternate forms for a total of twelve tests per recruit. Included in these were alternate forms of the NO and CS tests. Group 1 was administered all tests by computer and then by paper and pencil. Group 2 was administered all tests by paper and pencil and then by computer. Group 3 took all tests twice by paper and pencil.

A number of important results were obtained. First, as in the Greaud and Green study, the score means for CRT-1 were higher than they were in the paper-and-pencil version for the CS test. This was not the case, however, for the means for CRT-2. Regardless of which tests were presented first, the score means for CRT-1 were about 20 points higher than the means obtained by either paper and pencil or CRT-2. Score means for the latter two differed by between 1 and 2 points, with the higher means in the paper-and-pencil tests. In contrast, for the NO test, paper-and-pencil scores were significantly higher than the scores for either computer-administered version. This is the exact opposite of Greaud and Green's findings, albeit with a different dependent measure. Perhaps equally interesting, the magnitude of the differences between paper-and-pencil and computer versions of the NO test changed somewhat, depending on whether the computer test was given first or second. When the computer versions were administered first, paper-and-pencil means were almost 4 points higher than mean scores from CRT-1 (individually presented items), which in turn had higher mean scores (by about 6 points) than CRT-2 (multiple items per screen). When paper-and-pencil tests were administered first, the paper-and-pencil and CRT-1 mean scores differed by about a point. Both, however, were between 5 and 6 points higher than the mean obtained for CRT-2. Thus, there appears to be some sort of asymmetric practice effect. As stated earlier, such effects have important implications concerning appropriate designs for equating studies of automated and conventional tests. Specifically, single-group counterbalanced designs may not be a feasible data collection approach for equating studies with certain types of computer-administered tests.

Taken together, the results of both Greaud and Green (1986) and Kiely, Zara, and Weiss (1986) suggest that scores on computer-administered speed tests are not likely to be comparable with scores on the corresponding paper-and-pencil versions. Scores in coding speed tests like the ASVAB clerical speed test were higher when the test was administered by computer with items presented individually on the CRT. Results of the latter study suggest that scores more like those in the paper-and-pencil versions may be obtained when multiple items are pre-

sented on a single screen. The results were mixed with respect to arithmetic items like those in the NO test. In the study by Greaud and Green, presenting items individually resulted in higher scores, while in the study by Kiely, Zara, and Weiss, scores from the computer mode were somewhat lower, regardless of whether items were presented individually or in sets.

Tests Containing Graphics

The principal concern about tests that contain graphics is whether the resolution of the display medium used in a computer-administered version is sufficient to allow the examinee to correctly interpret, and extract the required information from, the pictures or diagrams. For timed tests or for speed tests, there is an additional concern whether the rate at which the material can be assimilated is slowed or facilitated, resulting in concomitant effects on score comparability.

Five studies that used tests containing graphics or both, are summarized in Table 10 (p. 24). For one of the studies (Reckase, Carlson, and Ackerman 1986), the tests were administered without a time limit, and no evidence for significant differences caused by mode of administration was obtained. In the remaining four studies, power tests were administered with a time limit. Two of these four report statistically significant differences in mean scores obtained under computer-administered and paper-and-pencil-administered conditions. However, both studies contain experimental design features that make their results somewhat unclear.

Reckase, Carlson, and Ackerman compared the item difficulty parameter estimates for the modified three-parameter IRT model obtained under paper-and-pencil and computer conditions for items from a radar technician training course. A total of 86 multiple-choice items were used to produce two different 43-item test forms, which were administered in a paper-and-pencil format. In addition, a series of 43-item test forms, generated by random selection from the total 86-item pool, were created and administered by the computer. The computer tests were on a Hazeltine terminal driven by an Ohio Scientific Challenger computer. Items were individually presented on a CRT, and examinees entered their answers through a computer keyboard. Details about opportunities to review or change answers are not given in the article. A total of 484 examinees in the radar technician course took one of the paper-and-pencil forms, while a second group of 427 examinees took the other paper-and-pencil form. The computer tests were taken by a total of 229 examinees.

Of the 86 items, sufficient data were obtained to allow comparison of parameters estimated from paper-and-pencil and computer administrations for only 37 items. Item parameter difficulty estimates were determined separately for each mode of test administration. If no mode-of-administration effects were present, parameter estimates from the two conditions should be linearly related. Reckase, Carlson, and Ackerman performed a number of analyses based on principal components analysis. The results suggest that if mode-of-administration effects were present, they were relatively small. However, the results of the study are difficult to interpret because both item order and item context effects were allowed to differ between the computer and the paper-and-pencil conditions. Effects associated with these factors may obscure possible mode-of-administration effects.

In the Wildgrube (1982) study (of the German Federal Armed Forces Aptitude Classification Battery discussed earlier), three of the tests—figural reasoning (20 items administered with a 10-minute time limit), mechanical ability (20 items administered with a 13-minute time limit), and electrotechnical (20 items administered with a 10-minute time limit)—contain varying amounts of figural graphics. Score means for all three tests were higher when they were administered by computer than when they were administered by paper and pencil, and the differences were statistically significant. All the mean differences between the two modes were approximately 1 point on a 20-point test. It is interesting to note that although they are timed, none of the tests is likely to be very speeded. Once again, however, the reader should be reminded that the Wildgrube results are difficult to interpret in that order effects may be confounded with mode-of-administration effects, because all computer tests were administered subsequent to the paper-and-pencil tests.

A better-controlled study, using similar tests, was that of Kiely, Zara, and Weiss (1986). Three of the six ASVAB tests studied—auto and shop information (AS), mechanical comprehension (MC), and electronics information (EI)—contain graphics in their problems. They were specifically included in the study to determine whether difficulties would arise related to the computer display of graphics. A single computer version of each test was produced. As described previously for two other ASVAB tests, 997 Air Force recruits were divided into three general experimental groups. Group 1 was administered all tests by computer and then by paper and pencil. Group 2 was administered all tests by paper and pencil and then by computer. Group 3 took all tests twice by paper and pencil.

No main effects for paper-and-pencil versions versus computer versions were detected for any of the three tests. However, for the MC and EI tests, there was a significant interaction between test order and presentation mode. For both tests the type of effect observed was similar. Paper-and-pencil scores tended to decrease slightly from the first to the second administration (about 0.2 points on 20- to 25-item tests), while mean scores on the computer-administered tests increased (between 0.6 and 1.4 points). As stated earlier,

such asymmetric practice/fatigue effects have important implications concerning tractable designs for equating conventional and automated tests.

Two studies compared scores from a paper-and-pencil version of the Hidden Figures Test (HFT) with scores from a computer-administered version. The HFT is intended to measure the field dependence/field independence construct. Examinees are presented with five sample figures, and their task is to indicate which of these figures is embedded in each of a number of complex patterns.

In the study by Jonassen (1986), scores from a computer version of part of form CF1 of the HFT were compared with a paper-and-pencil version of a different part of the same form. The first part (patterns 1–16) of the form was administered in the standard paper-and-pencil format while the second part (patterns 17–32) was administered by computer. The computer-administered version was presented on an Apple microcomputer with a 12-inch monochrome screen. All figures and patterns in the microcomputer version were identical in proportion and nearly identical in size to those in the paper-and-pencil version. The instructions, timing (12 minutes for each part of the test), and scoring were the same in both testing conditions. Again, however, several important differences existed between the paper-and-pencil and the computer versions.

In the paper-and-pencil test all 16 patterns were displayed on two pages in three rows (that is, a total of six rows, four containing 3 patterns and two containing 2 patterns). Beneath each pattern were the letters *A* through *E*, which corresponded to the five sample figures. The examinee circled the letter corresponding to the sample figure he or she thought was embedded in each of the complex patterns. In the computer versions the patterns were displayed a row at a time, and the subjects had to flip between rows by using the left and right arrow keys on the microcomputer keyboard. When the examinees were ready to answer, they first typed an *A*, indicating their readiness. Next, the examinees were required to type a number from 1 to 16, indicating which pattern they wanted to answer. Finally, they were to type the letter corresponding to the sample figure they thought was embedded in the complex pattern. Examinee answers were then displayed, and the examinees were asked to verify the correctness of the response or to change the answer.

The participants in the study were 58 upper-level undergraduate and graduate education and library science students. All had some experience with microcomputers. One group (N = 38) received the paper-and-pencil portion of the test, followed immediately by the computer portion. A second group (N = 20) was tested in the reverse order.

Score means for the paper-and-pencil test were 1.33 points higher (out of 16 points) than the mean scores for the computer portion. Although this difference is statistically significant, it is difficult to interpret since different half-tests were administered in each of the two administration modes. Differences could at least partially be attributable to differences in difficulty between the half-tests. When analyzed separately for the group who took the paper-and-pencil portion first, the means were identical in the two modes. However, score means were dramatically lower on the computer portion for the group who took that portion first (2.33 versus 5.20 out of 16 items). As in previously discussed studies, these kinds of results may have important implications for future equating efforts with computerized tests.

In a second HFT study, by Jacobs, Byrd, and High (1985), the test was adapted to be administered by the PLATO system. All the figures and patterns in the computer version were identical in proportion and nearly identical in size to those in the paper-and-pencil version. The instructions, timing, and scoring were the same in both testing conditions. Again, however, the items were displayed differently. In the paper-and-pencil test, sample figures were displayed in a single row. The complex patterns were displayed on pages, each of which contained three rows of figures. Beneath each pattern were the letters *A* through *E*, which corresponded to the five sample figures. In the computer version, the row of sample figures and a single complex pattern were displayed simultaneously on a single computer screen.

Forty undergraduate students took the HFT twice, once in the paper-and-pencil format and once by computer. All retesting was done immediately. The order of administration was counterbalanced across administrations. Score means did not differ significantly across the two modes. However, when the test was administered in the paper-and-pencil format, means were over 2 points higher (out of 32 points) than they were when the test was administered by computer (15.97 versus 13.85). This corresponds to an effect size of 0.25. In addition, the correlation between scores in the two conditions was only 0.57. This is surprisingly low since the same items were immediately readministered in different presentation modes.

In sum, the evidence for timed power tests is mixed concerning whether the display of graphics has an effect on the relative difficulty of computer and paper-and-pencil versions of a test. One study that found significant mode-of-administration effects (Wildgrube 1982) is difficult to interpret because of possible confounding with order effects. A second study using similar tests (Kiely, Zara, and Weiss 1986) found no mode-of-administration main effects; however, a significant test mode by order interaction was observed for two of the three tests in the study. Two studies (Jacobs, Byrd, and

High 1985; Jonassen 1986) found computer versions of the HFT more difficult than were paper-and-pencil versions. However, only in the latter study were the differences statistically significant. Unfortunately, in this same study, differences in difficulty between the half-tests administered in the paper-and-pencil and computer versions of the HFT may be confounded with mode-of-administration effects. Considerably more research will be required to determine exactly what factors influence the presence of mode-of-administration effects for timed power tests with graphics.

Tests Containing Items Based on Reading Passages

Many tests contain items based on a reading passage or a set of figures. It is often impossible to simultaneously display an item along with the complete passage or set of figures to which the item refers.

A number of factors may contribute to mode-of-administration effects for computer tests containing items based on reading passages. First, examinees must be concerned with routing themselves through a number of related screens to obtain information necessary to complete an item; this may be distracting or may make a given task more cognitively demanding. The result could be that computer versions appear more difficult than their paper-and-pencil counterparts. Second, particularly for timed tests, the number of keystrokes involved and the time required for CRT display may cause computer versions to be more speeded and therefore to appear more difficult.

Two studies that used tests containing items based on reading passages are summarized in Table 11. In both studies statistically significant differences were obtained.

An early attempt to adapt a paper-and-pencil test, containing items based on reading passages, for computer administration is described by Feurzeig and Jones (1970). They administered two equivalent forms of the Triggs Diagnostic Reading Test to a group of 61 eighth-grade students. The test consists of a single long passage (about 250 lines) followed by 20 multiple-choice questions based on the passage. Each student received one form of the test in the standard paper-and-pencil mode, and the other form was administered by computer.

In the computer-administered version, the reading material was presented in segments of approximately three lines on a TTY. The students controlled the pacing of the presentation by striking a teletype key whenever they were ready to view the next segment of text. They controlled the presentation of multiple-choice test items in the same fashion. The computer version was set up so that the students could skip items and go back and change previous answers. Approximately half the subjects received the paper-and-pencil test first, while the other half received the computer version first. No time limits were imposed in either administration.

It should be noted that having the material presented on a TTY is quite different from having the same setup on a CRT. As subsequent segments of text are produced, the previous segments are still visible to the examinee and can be viewed as desired. In this respect there is no difference between a TTY computer version and a paper-and-pencil version. In contrast, when the number of segments presented on a CRT exceeds the capacity of the screen, previously presented segments must be either written over or scrolled off the screen. These segments are no longer immediately available and must be sought if the information they contain is required. Thus, the main difference between the TTY version of the diagnostic reading test developed by Feurzeig and Jones and the paper-and-pencil version resides solely in the paced presentation (three lines at a time) of the reading passage.

Despite the similarities between the TTY and paper-and-pencil versions, results indicated that administering the test by computer did have an effect on the distribution of scores. The scores on the computer test ranged from 6 to 17, with a mean of 12.1 and a standard deviation of 2.9. The scores on the paper-and-pencil test were, on the average, slightly lower and somewhat more variable. They ranged from 2 to 18, with a mean of 11.0 and a standard deviation of 4.3. The difference in mean scores is not statistically significant. However, the difference in standard deviations is significant (at the 0.01 level). Feurzeig and Jones attribute these differences to a relatively higher level of performance on the computer test by low-ability students. They speculate that the attention-getting properties of having to interact with the computer and the control of the pacing of the material are especially helpful to low-ability students. However, different test forms were administered in the two conditions, and at least some of the score differences may be due to differences in the psychometric properties of the test forms rather than to mode-of-administration effects.

One of the tests studied by Kiely, Zara, and Weiss (1986), the ASVAB paragraph comprehension (PC), requires subjects to answer sets of questions based on reading passages that are too long to be displayed on a single screen. Three different computer versions of the PC test were developed. In the first, CRT-1, the paragraph under consideration could be scrolled as one question at a time appeared under it on the screen. Examinees were unable to refer to previous paragraphs or to previous questions about the current paragraph. In the second computer version, CRT-2, the paragraph and all related questions were presented in a single scrollable field. Examinees could move back and forth from the current paragraph to its related questions but could not review previous paragraphs or the questions related to

Table 10. Summary of Studies Using Tests Consisting of Single-Screen Items with Pictorial or Figural Information

Study	Test	Type	Experimental Design				Results	
			Mode Factor	Controlled for: Order	Form	Subjects	Significant Differences	Statistics Affected
Reckase, Carlson, & Ackerman (1986)	Radar technician test	Untimed	B*	—	—	N	—	—
Wildgrube (1982)	Figural reasoning	Timed power	W†	N‡	Y	—	Y	Mean
	Mechanical ability	Timed power					Y	Mean
	Electrotechnical	Timed power					Y	Mean
Jonassen (1986)	Hidden Figures Test	Timed power	W	Y§	N	—	Y	Mean
Kiely, Zara, & Weiss (1986)	ASVAB auto & shop information	Timed power	W	Y	Y	—	N	—
	ASVAB mechanical comprehension	Timed power					Y	Mean
	ASVAB electronics information	Timed power					Y	Mean
Jacobs, Byrd, & High (1985)	Hidden Figures Test	Timed power	W	Y	Y	—	N	—

*B = between-subject variable.

†W = within-subject variable.

‡N = no

§Y = yes

Table 11. Summary of Studies Using Tests Containing Multiscreen Items

Study	Test	Type	Experimental Design				Results	
			Mode Factor	Controlled for: Order	Form	Subjects	Significant Differences	Statistics Affected
Feurzeig & Jones (1970)	Triggs Diagnostic Reading Test	Untimed	W	Y	Y	—	Y	Standard deviations
Kiely, Zara, & Weiss (1986)	ASVAB paragraph comprehension	Timed power	W	Y	Y	—	Y	Mean

them. In the third computer version, CRT-3, paragraphs and questions appeared in separate scrollable fields. An added feature of CRT-3 was the inclusion of a display at the top of the screen that allowed examinees to monitor their progress through the test and to change answers to all previous PC questions.

Data analyses revealed a number of statistically significant effects. Overall, the mean scores on the paper-and-pencil version were 2.6 points higher (on a 15-item test) than they were across the three computer-administered versions. The mean scores for CRT-1 (paragraph scrolling with only a fixed item displayed) and CRT-3 (independent scrolling of paragraphs and items) did not differ. However, mean scores for both were significantly higher (between 1.5 and 2 points on a 15-item test) than the mean score for CRT-2 (scrolling both paragraphs and items in a single field). It should be noted that in both CRT-1 and CRT-3 the examinee can view different sections of the reading passage with the item in view. This is not possible for CRT-2 because all information is in a single field; if the field is occupied by an item, it cannot contain a reading passage, and vice versa.

In sum, the study most relevant to current attempts to develop computer-administered versions of tests containing items based on reading passages (Kiely, Zara, and Weiss) reported significant mode-of-administration effects. Mean scores were somewhat lower when the tests were administered by computer. Despite a number of features designed to make them flexible to work with, all three computer versions produced by Kiely, Zara, and Weiss resulted in scores that were lower than those in the paper-and-pencil version. However, equally clearly, being able to scroll through the reading passage while an item is concurrently displayed is a desirable feature to be incorporated in a computer test that contains reading passages.

CONCLUSIONS

Many studies have been reviewed, some in detail. The studies have used automated versions of different types of tests, different approaches to presenting test items, and different devices for recording examinee responses. While several studies present evidence that scores from automated versions are more or less equivalent to their conventional counterparts, an equally impressive number present evidence to the contrary. More important, mode-of-administration effects are observed not only for tests containing complex multiscreen item types (like reading comprehension tests) and for speed tests, where such effects might be expected, but also for tests with relatively simple item formats (like personality inventories).

On the basis of our review, evidence concerning a number of factors related to the implementation of automated testing procedures has been identified, which suggests these factors may have an effect on score equivalence. We have highlighted what we believe are important conclusions in section summaries throughout the report. Several points, however, are particularly important and are worth repeating.

1. The rate at which test-takers omit items in an automated test may differ from the rate at which items are omitted in a conventional presentation. Particular care will need to be exercised in producing automated versions of tests that are formula-scored.
2. There is evidence that scores on tests from automated versions of personality inventories such as the MMPI are lower than scores obtained in the conventional testing format. These differences may be due in part to differing omit rates, as described above, but some of the differences may be caused by other factors.
3. Scores from automated versions of speed tests will most likely not be comparable with paper-and-pencil versions. This is probably because of the greater ease with which answers can be indicated when the test is administered by computer.
4. Studies using the Hidden Figures Test have demonstrated that the presentation of figural and pictorial information can have an effect on score equivalence, making the test more difficult. However, in studies with three ASVAB tests, effects were not found. Further research is needed to determine what factors are related to the effect graphic information may have on scores.
5. Based primarily on the results of a single study with the ASVAB, tests containing items like those in reading comprehension tests may be more difficult when presented on a CRT.
6. In several instances, groups that were administered an automated version of a test followed by a conventional version exhibited substantial increases in test scores, while comparable groups given the tests in the opposite order showed little or no increase. The possibility of such asymmetric practice effects may make it wise to avoid conducting equating studies based on single-group counterbalanced designs.

All the above conclusions are tentative and are based on a review of the results of the studies conducted to date. Many, if not most, of the studies contain rather small sample sizes, which raises concerns as to the generalizability of the results. In addition, the results from several of the studies are difficult to interpret because factors such as practice or fatigue and differences in form

difficulty are confounded with the mode of test administration. Thus, the conclusions listed will need to be amended and qualified as additional evidence accumulates. *Despite the tentative nature of our conclusions, it is clear that test publishers need to perform separate equating and/or norming studies when computer-administered versions of standardized tests are introduced.*

REFERENCES

American Psychological Association. 1986. *Guidelines for computer-based tests and interpretations.* Washington, D.C.: American Psychological Association.

Anastasi, A. 1976. *Psychological testing.* 4th ed. New York: Macmillan.

Biskin, B. H., and R. L. Kolotkin. 1977. Effects of computerized-administration on scores on the Minnesota Multiphasic Personality Inventory. *Applied Psychological Measurement* 1(4): 543–49.

Butcher, J. N. 1987. *Computerized psychological assessment.* New York: Basic Books.

Calvert, E. J., and R. C. Waterfall. 1982. A comparison of conventional and automated administration of Raven's Standard Progressive Matrices. *International Journal of Man-Machine Studies* 17:305–10.

Curtis, H. A., and R. P. Kropp. 1961. A comparison of scores obtained by administering a test normally and visually. *Journal of Experimental Education* 29(3): 249–60.

Denner, S. 1977. Automated psychological testing: A review. *British Journal of Social and Clinical Psychology* 16:175–79.

Dunn, T. G., R. E. Lushene, and H. F. O'Neil. 1972. Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology* 39:381–87.

Elwood, D. L. 1969. Automation of psychological testing. *American Psychologist* 24:287–89.

Elwood, D. L. 1972a. Test retest reliability and cost analyses of automated and face-to-face intelligence testing. *International Journal of Man-Machine Studies* 4:1–22.

Elwood, D. L. 1972b. Automated WAIS testing correlated with face-to-face WAIS testing: A validity study. *International Journal of Man-Machine Studies* 4:129–37.

Elwood, D. L. 1972c. Automated versus face-to-face intelligence testing: Comparison of test-retest reliabilities. *International Journal of Man-Machine Studies* 4:363–69.

Elwood, D. L. 1972d. Validity of an automated measure of intelligence in borderline retarded subjects. *American Journal of Mental Deficiency* 77(1): 90–94.

Feurzeig, W., and G. Jones. 1970. *Reevaluating low achievers with computer-administered tests.* Boston: Bolt, Beranek, and Newman.

Greaud, V. A., and B. F. Green. 1986. Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement* 10(1): 23–24.

Green, B. F., R. D. Bock, L. G. Humphreys, R. L. Linn, and M. D. Reckase. 1984. Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement* 21:347–59.

Hedl, J. J., H. F. O'Neil, and D. N. Hansen. 1971. Computer-based intelligence testing. Paper presented at annual meeting of the American Educational Research Association, February, New York, N.Y.

Hitti, F. J., R. L. Riffer, and E. R. Stuckless. 1971. Computer-managed testing: A feasibility study with deaf students. Rochester Institute of Technology, National Technical Institute for the Deaf, Rochester, N.Y. Typescript.

Hoffman, K. I., and G. D. Lundberg. 1976. A comparison of computer-monitored group tests with paper-and-pencil tests. *Educational and Psychological Measurement* 36:791–809.

Jacobs, R. L., D. M. Byrd, and W. R. High. 1985. Computerized testing: The Hidden Figures Test. *Journal of Educational Computing Research* 1(2): 173–77.

Johnson, D. F., and W. L. Mihal. 1973. Performance of blacks and whites in computerized versus manual testing environments. *American Psychologist* 28:694–99.

Jonassen, D. H. 1986. Effects of microcomputer display on a perceptual/cognitive task. Paper presented at annual meeting of the Association for Educational Communications and Technology, January, Las Vegas.

Katz, L., and J. T. Dalby. 1981. Computer-assisted and traditional psychological assessment of elementary-school-aged children. *Contemporary Educational Psychology* 6:314–22.

Kiely, G. L., A. R. Zara, and D. J. Weiss. 1986. Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests. Research Report No. AFHRL-TP-86-13. Air Force Human Resources Laboratory, Brooks Air Force Base, Tex.

Klinge, V., and T. Rodziewicz. 1976. Automated and manual intelligence testing of the Peabody Picture Vocabulary Test on a spycgiatric adolescent population. *International Journal of Man-Machine Studies* 8:243–46.

Knights, R. M., D. H. Richardson, and L. R. McNarry. 1973. Automated versus clinical administration of the Peabody Picture Vocabulary Test and the Coloured Progressive Matrices. *American Journal of Mental Deficiency* 78(2): 223–25.

Koson, D., C. Kitchen, M. Kochen, and D. Stodolosky. 1970. Psychological testing by computer: Effect of response bias. *Educational and Psychological Measurement* 30:803–10.

Leary, L. F., and N. J. Dorans. 1985. Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research* 55(3): 387–413.

Lee, J., and L. Hopkins. 1985. The effects of training on computerized aptitude test performance and anxiety. Paper presented at annual meeting of the Eastern Psychological Association, April, Boston.

Lee, J., K. E. Moreno, and J. B. Sympson. 1986. The effects of mode of test administration on test performance. *Educational and Psychological Measurement* 46:467–74.

Lukin, M. E., T. Dowd, B. S. Plake, and R. G. Kraft. 1985. Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior* 1:49–58.

Lushene, R., H. O'Neil, and T. Dunn. 1974. Equivalent validity of a completely computerized MMPI. *Journal of Personality Assessment* 13:407–12.

McBride, J. R., and D. J. Weiss. 1974. *A word knowledge item pool for adaptive ability measurement.* Research Report No. 74-2. Minneapolis-St. Paul: University of Minnesota.

Olsen, J. B., D. D. Maynes, D. Slawson, and K. Ho. 1986. Comparison of equating of paper-administered, computer-administered, and computerized adaptive tests of achievement. Paper presented at annual meeting of the American Educational Research Association, April, San Francisco.

Overton, G. N., and K. G. Scott. 1972. Automated and manual intelligence testing: Data on parallel forms of the Peabody Picture Vocabulary Test. *American Journal of Mental Deficiency* 76(6): 639–43.

Parks, B. T., D. E. Mead, and B. L. Johnson. 1985. Validation of a computer-administered marital adjustment test. *Journal of Marital and Family Therapy* 11(2): 207–10.

Reckase, M. D., J. E. Carlson, and T. A. Ackerman. 1986. The effect of computer presentation on the difficulty of test items. Paper presented at annual meeting of the American Psychological Association, August, Washington, D.C.

Sachar, J. D., and J. D. Fletcher. 1978. Administering paper-and-pencil tests by computer, or the medium is not always the message. In *Proceedings of the 1977 computerized adaptive testing conference,* ed. D. J. Weiss. Minneapolis: University of Minnesota.

Scissons, E. H. 1976. Computer administration of the California Psychological Inventory. *Measurement and Evaluation in Guidance* 9(1): 22–25.

Space, L. 1981. The computer as psychometrician. *Behavior Research Methods and Instrumentation* 13(4): 595–606.

White, D. M., C. B. Clements, and R. D. Fowler. 1985. A comparison of computer administration with standard administration of the MMPI. *Computers in Human Behavior* 1:153–62.

Wildgrube, W. 1982. Computerized testing in the German Federal Armed Forces: Empirical approaches. In *Item Response Theory and Computerized Adaptive Testing Conference Proceedings,* ed. D. J. Weiss, 353–59. Minneapolis: University of Minnesota.

Wise, L. A., and S. L. Wise. 1986. Comparison of computer-administered and paper-administered achievement tests with elementary school children. Department of Educational Psychology, University of Nebraska, Lincoln. Typescript.

Wise, S. L., L. L. Boettcher, A. L. Harvey, and B. S. Plake. 1987. Computer-based testing versus paper-and-pencil testing: Effects of computer anxiety and computer experience. Paper presented at annual meeting of the American Educational Research Association, April, Washington, D.C.

218111